Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Challenges with evaluating education policy using panel data during and after the COVID-19 pandemic

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Stuart, Elizabeth A.	John Hopkins Bloomberg School of Public Hea	
Feller, Avi	University of California Berkley	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) July 30 2021

Check type of content being submitted and complete one of the following in the box below:

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Journal of Research on Educational Effectiveness 14,3

DOI or URL to published work (if available) https://doi.org/10.1080/19345747.2021.1938316

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** _____ through **[Grant number]** _____ to **Institution]** _____.The opinions expressed are those of the authors and do not represent views of the **[Office name]** _____ or the U.S. Department of Education.

Challenges with evaluating education policy using panel data during and after the COVID-19 pandemic[1]

Avi Feller and Elizabeth A. Stuart

https://doi.org/10.1080/19345747.2021.1938316

Abstract

Panel data methods, which include difference-in-differences and comparative interrupted time series, have become increasingly com- mon in education policy research. The key idea is to use variation across time and space (e.g., school districts) to estimate the effects of policy or programmatic changes that happen in some localities but not others. In this commentary we highlight some specific challenges for panel or longitudinal studies of K-12 education interventions during and following the COVID-19 pandemic. Our goal is to help researchers think through the underlying issues and assumptions, and to help consumers of those studies assess their validity.

[1] We would like to thank Kelly Hallberg, Luke Miratrix, and Jesse Rothstein for helpful comments and discussion. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Dr. Stuart's time was supported by a Johns Hopkins University Discovery Award and NIMH award P50MH115842.

The COVID-19 pandemic has upended countless lives and transformed teaching and learning across the world. The pandemic has also upended education researchers' ability to evaluate those changes --- as well as unrelated interventions from before the pandemic.

In this short note, we highlight some specific challenges for panel or longitudinal studies of K-12 education interventions during and following the COVID-19 pandemic. Our goal is to help researchers think through the underlying issues and assumptions, and to help consumers of those studies assess their validity. While we focus on the implications for education research, other papers discuss policy evaluation in this context more broadly (e.g., Bacher-Hicks and Goodman, 2020; Ben-Michael, Feller, and Stuart, 2020; Goodman-Bacon and Marcus, 2020; Haber et al., 2020).

Panel data methods, which include difference-in-differences and comparative interrupted time series, have become increasingly common in education policy research. The key idea is to use variation across time and space (e.g., school districts) to estimate the effects of policy or programmatic changes that happen in some localities but not others. These designs are strongest when: outcomes are relatively stable over time; the comparison localities (say states, districts, or schools) provide a good proxy for what would have happened in the intervention localities had they not changed their policies or programs; and there were no other co-occurring policy or programmatic changes in the intervention localities. A wide variety of statistical methods can be used to estimate effects in this longitudinal data context; for summaries of these methods in education research, see Bloom (2003), Jacob et al. (2016), and Hallberg et al. (2018).

While these methods can be powerful tools for education research in typical times, our outlook on the use of panel data modeling in the context of the COVID-19 pandemic is rather pessimistic. Substantively, this is difficult because the combination of intervention, measurement, and setting are often highly idiosyncratic to the COVID-19 context. Statistically, this is difficult because of the myriad sets of policies and changes students and schools experienced this year, and because outcome trajectories during the pandemic are such a departure from the recent past. In addition, there is the question of whether and how lessons from COVID-specific interventions can inform post-COVID policies.

Some new challenges with panel data methods for education research

We focus on analyses using aggregate data available over time, such as test scores, attendance rates, COVID case rates (e.g., Harris et al., 2021), or even area-level mental health measures for children or their parents (e.g., Davis et al., 2020). Much of what we discuss, though, would be relevant for settings where there is individual-level data available within the units under study. In some examples, COVID-related changes are central to the analysis, such as in investigating the link between in-person schooling and COVID cases or hospitalizations (Harris et al., 2021; Lessler et al., 2021). In other examples, COVID-related changes are not a

central part of the analysis but rather a substantial problem to be addressed, such as studying the use of a particular reading curriculum on reading assessment scores. For both types of examples, the goal is to use the panel data structure to estimate the impact of an education policy change---moving instruction onsite or adopting a specific curriculum---while recognizing that districts and states do not make these changes randomly.

We start with important conceptual and measurement concerns and then turn to estimation challenges inherent in the use of panel data methods:

- **Defining the policy and effect estimate of interest.** The first challenge is to precisely define the research question and the intervention of interest: what is the precise policy or program under study? While this may sound obvious, it can be challenging in practice. One example of the challenge is in evaluating whether in-person schooling is associated with increased COVID-19 rates in the community, or better social and emotional or academic outcomes for students. The specifics of "in-person instruction" in districts in which it is offered vary widely, and change rapidly over time in many locations. Thus, clearly defining and measuring the intervention is both crucial and complex (Lupton-Smith et al., 2021).

- **Isolating the policy of interest.** Even with a well-defined intervention, it can be difficult to isolate settings in which only the policy change of interest occurs (Bacher-Hicks and Goodman, 2020). Especially during 2020, students, families, teachers, and schools were subject to never-ending policy changes. For example, changes to online instruction were accompanied by stay at home orders and increased physical distancing, making it challenging to disentangle, say, the effects of virtual instruction on students' mental health from these other changes to their routines and experiences. As another example, a new reading curriculum may have been rolled out during the 2020-2021 academic year, but accompanied by some periods of time with virtual rather than in-person instruction.

- **Measurement.** Of course what is also needed are measures of the outcome of interest over time, before and after the policy change of interest. This too can be a challenge. Many states and school districts have paused annual assessments and other education data collection (National Academy of Education, 2021). When the data are available, there are important open questions about measurement equivalence pre- and post-pandemic, and how to appropriately consider changes in measures over time (Boyer, 2021). Thus, even before considering issues related to causal effect estimation, researchers should consider whether and how they might use available assessment data. And, as we discuss below, these questions also have implications for the parallel trends assumption common with panel data methods.

- **Missing data.** These questions are even more challenging for assessments in 2020 and 2021 that are simply missing (Connors et al., 2021). True always --- but especially during the pandemic --- not all missingness is created equal. For example, missingness due to

the massive disruption in Spring 2020 is widespread, while missingness during Fall 2020 and Spring 2021 likely varies dramatically based on district- or state-specific choices, leading to different patterns of missingness. Moreover, student-level missingness likely changed over the course of the pandemic, reflecting the compounding disparities. Importantly, student-level missingness remains a concern even when considering interventions at the aggregate level, since changing student composition complicates comparisons over time.

After considering these fundamental questions, we now turn to estimation challenges more specific to using panel data methods, namely around the key underlying assumptions and selecting valid comparisons:

- **Violating parallel counterfactual trends.** The primary estimation hurdle to applying standard panel data methods during the COVID-19 pandemic is that the key assumptions required to obtain accurate estimates of policy effects are unlikely to hold. In particular, panel data approaches typically assume versions of *parallel counterfactual trends*: in the absence of the policy change, the relationship between outcomes for the treated and comparison localities would remain stable over time (including into the post-policy time period). For instance, CITS starts with a model of the time series trend for the treated units (a single "interrupted time series") and then uses a comparable model for the control units to adjust the estimate. The key assumption is that---if the policy change had never occurred---this modeled relationship would continue into the post-policy period (Hallberg et al., 2018).

  Assumptions like this can be difficult to justify even in normal times, but are especially challenging during COVID. Intuitively, the key assumption is that areas with similar pre-treatment trends had the same reaction to COVID---not including the policy under investigation---that also did not affect the scale of the outcome. This seems difficult to justify in typical education settings: it is hard to forecast how quickly outcomes of interest (such as assessments) will start to return to pre-pandemic levels, what the shape of that evolution will be, and whether it might vary considerably across localities, especially given the wide variation in how schools operated during the pandemic.

  Moreover, since these are assumptions about *counterfactual* trends (e.g., the trend that would have happened in the treated units had they not actually implemented the policy), researchers can only assess them indirectly. Common practice is to examine the *pre-treatment* relationship between treated and comparison localities, and, if the relationship is stable, to argue that it will persist through the post-treatment period (Haber et al., 2020). The context of the COVID-19 pandemic, and especially the measurement challenges referenced above, makes it even more difficult to use pre-treatment estimates to reason about counterfactual assumptions, meaning that standard methods for interrogating those assumptions are far less useful.

- **Finding valid comparisons in extraordinary times.** A related question is how to construct reasonable comparison groups for CITS models. How best to find comparisons for this approach remains an open question in general (Hallberg et al., 2018), and the questions and concerns are magnified during and after the COVID-19 pandemic. For example, the decision to move instruction online likely depends on local COVID-19 case loads, political factors, and hospital capacity, among other factors. These are not typically included as possible confounders in, say, the decision to adopt a new reading curriculum, but the schooling modality would be crucial information for assessing the plausibility of a comparison locality.

  In addition, the pandemic has exacerbated existing disparities across well-established racial and socio-economic lines. Thus, in some settings it might be possible that matching along these dimensions might also be good proxies for pandemic responses more generally --- and that failing to match on these dimensions may lead to serious challenges in the validity of the comparison localities.[2]

- **Generalizability.** Finally, even if researchers are able to address all of these statistical and conceptual issues, they must grapple with the extent to which effect estimates from the pandemic can inform post-pandemic decisions, when the context will be yet again fundamentally different. For instance, if an intervention is found to improve student engagement with course materials during online instruction from home, how likely is it that a similar result would be seen once students are back in their classrooms?

Of course, these are just some of the many challenges with panel data methods affected by COVID-19 (Goodman-Bacon and Marcus, 2020). For example, spillover between units --- a common challenge with students interacting inside a classroom and thus familiar to education researchers for years --- is now a concern for the spread of infectious diseases across neighboring districts. In addition, many COVID-19 studies have to deal with the wide number of policies that were implemented around the same time in many places around the country and world, making it hard to disentangle the effects of the different component policies. We note, too, that these challenges are not unique to education research. Haber et al. (2021) provide a review of the strength of evidence from existing panel data studies examining COVID-19 outcomes (with a broad range of exposures of interest, including mask-wearing mandates and stay-at-home orders), and found that few met minimal criteria for rigorous policy evaluations.

Looking ahead

While there are no easy solutions to these challenges, we do have some suggestions for how to make progress towards generating rigorous evidence about the effects of education-related policies and programs during and after the COVID-19 pandemic.

[2] We thank Kelly Hallberg for this point.

- **Incorporate non-traditional quantitative data sources.** In the absence of traditional assessments, researchers have started to embrace non-traditional data sources that may have more consistent measurement and meaning before, during, and after the pandemic. For example, Chetty et al. (2020) use data from the online learning platform Zearn to assess learning loss for a sub-population of students. Bacher-Hicks et al. (2020) examine search trends regarding online educational tools. Other online educational technology tools such as iReady and Khan Academy may provide data to model trends, rather than relying on annual standardized test scores. This will necessarily involve care and attention to understanding the limitations of these sources. In particular, we must understand which populations are underrepresented or entirely missing from these data sets, and, as we discuss above, whether the composition of those students with observed data changes over time.

- **Increase quantitative data collection on the situation on the ground.** We need to collect data to understand what has been happening on the ground in school districts during the pandemic. Understanding which districts offered onsite instruction, what that looked like, and how many students participated, will be crucial to help interpret key outcomes during and after the pandemic. There are some data collection efforts underway that touch on these ideas, such as the US Census Bureau Household Pulse Survey, the AIR National Survey of Public Education's Response to COVID-19, and a few data sources tracking school opening decisions (Lupton-Smith et al., 2021). However, they are fairly limited in their scope and scale, giving, for example, snapshots of what was happening in fairly narrow time windows, or broader data but without much details about instruction. This may be another area in which administrative data, such as virtual attendance or engagement data, or ideally onsite attendance data, will be useful. A potentially novel source of this data is cell phone tracking data, as aggregated by Parolin and Lee (2021).

- **Incorporate qualitative data sources and case studies.** While qualitative evidence and case studies should always be part of a comprehensive analysis, they can play especially important roles in understanding what was happening "on the ground" for students, schools, and districts during the pandemic. Even just trying to describe school district policies around the mode of instruction requires extensive qualitative work to understand the sheer range of variations, and generating a body of case studies and qualitative evidence can help round out the (messy) pictures that might be obtained from quantitative policy evaluations. Related efforts are underway in studies of opioid policy, and could serve as a model (McGinty et al., 2021)

- **Pay attention to subgroups.** Many panel data analyses use data measured only at a full group level (e.g., all 3rd graders in a school or district). However, the pandemic is not influencing all populations the same. Looking at only school average test scores may miss serious implications for particular subgroups of students; care should be taken to collect data on important subgroups and study their outcomes in particular. Moreover,

looking at differences between these subgroups *within* the same context can be valuable, and even a particularly compelling comparison, given that the groups experienced the same contexts before and during the pandemic.

We hope that the coming years will include carefully done and rigorous research that helps us learn about the implications of the COVID-19 pandemic on students, teachers, and the educational system in general, as well as which interventions may have helped ameliorate any negative consequences. However, based on the current proliferation of COVID-related panel data papers (Haber et al., 2021), education researchers will likely read or review many studies with weak designs. For these, we hope that the bullet points above can provide a rough checklist for assessing study strength. For example, does the study clearly define the treatment and quantity of interest? Does it address possible measurement changes? How plausible is the relevant parallel trends assumption? (See Haber et al., 2020 for related checklists.) Hopefully these questions can steer readers towards more compelling studies.

Finally, while progress on answering these research questions is important, we must also "be clear about what is knowable" (Goodman-Bacon and Marcus, 2020, p. 156). Given data constraints and a rapidly changing world, we might never have comprehensive answers to many of the key empirical questions in education in 2020 and 2021. Identifying the best data and methods to answer these questions will be crucial, as will having the humility to understand when the data and context are too complex for reliable results.

References

Bacher-Hicks, A. & Goodman, J. (2020). The Covid-19 Pandemic Is A Lousy Natural Experiment for Estimating the Effects of Education Policy Changes. Working paper.

Bacher-Hicks, A., Goodman, J., & Mulhern, C. (2020). *Inequality in Household Adaptation to Schooling Shocks: Covid-Induced Online Learning Engagement in Real Time* (No. 27555). National Bureau of Economic Research. **DOI** 10.3386/w27555

Ben-Michael, E., Feller, A., & Stuart, E. A. (In press). A trial emulation approach for policy evaluations with group-level longitudinal data. *Epidemiology*.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*(1), 3-49.

Boyer, M. (2021). Focus, Fix, Fit: Understanding the Meaning of 2021 Test Scores: Finding a Path Forward with Existing Tools and Procedures. National Center for the Improvement of Educational Assessment. Available at: https://www.nciea.org/blog/state-testing/focus-fix-fit-understanding-meaning-2021-test-scores .

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., & The Opportunity Insights Team. (2020). *The economic impacts of COVID-19: Evidence from a new public database built using private sector data* (No. w27431). National Bureau of Economic Research.

Connors, M. C., Easton, J. Q., Ehrlich, S. B., Feller, A., Francis, J., & Kabourek, S. E. (2021) Missing Data Due to COVID: A Case Study of Pre-K in Chicago. Working paper.

Davis, C. R., Grooms, J., Ortega, A., Rubalcaba, J. A. A., & Vargas, E. (2020). Distance Learning and Parental Mental Health During COVID-19. *Educational Researcher*, 0013189X20978806.

Goodman-Bacon, A., & Marcus, J. (2020, June). Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. In *Survey Research Methods* (Vol. 14, No. 2, pp. 153-158).

Haber, N. A., Clarke-Deelder, E., Salomon, J. A., Feller, A., & Stuart, E. A. (2020). Policy evaluation in COVID-19: A graphical guide to common design issues. *arXiv preprint arXiv:2009.01940*.

Haber, N. A., Clarke-Deelder, E., Feller, A., Smith, E. R., Salomon, J., MacCormack-Gelles, B., ... & Stuart, E. A. (2021). Problems with Evidence Assessment in COVID-19 Health Policy Impact Evaluation (PEACHPIE): A systematic strength of methods review. *medRxiv*.

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, *47*(5), 295-306.

Harris, D. N., Ziedan, E., & Hassig, S. (2021) The Effects of School Reopenings on COVID-19 Hospitalizations. National Center for Research on Education Access and Choice (REACH) working paper.

Jacob, R., Somers, M. A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, *40*(3), 167-198.

Lessler, J., Grabowski, M.K., Grantz, K.H., Badillo-Goicoechea, E., Metcalf, C.J.E., Lupton-Smith, C., Azman, A., and Stuart, E.A. (2021). Household COVID-19 risk and in-person schooling. *Science*. Published online 29 April 2021.

Lupton-Smith, C., Badillo Goicoechea, E., Lessler, J., Grabowski, K., and Stuart, E.A. (2021). Measuring school openings and closings during the COVID-19 pandemic: Concordance across different data sources. https://arxiv.org/abs/2103.13296.

McGinty, E.E., Tormohlen, K.N., Barry, C.L., Bicket, M.C., Rutkow, L., and Stuart, E.A. (2021). Mixed-methods study of how implementation of U.S. state medical cannabis laws affects treatment of chronic non-cancer pain and adverse opioid outcomes. *Implementation Science* 16: 2.

National Academy of Education. (2021). Educational assessments in the COVID-19 era and beyond. Washington, DC: Author.